Research in Astron. Astrophys. Vol.0 (20xx) No.0, 000–000
http://www.raa-journal.org *http://www.iop.org/journals/raa*
(LATEX: Alwayshappy-Jia.tex; printed on July 20, 2023; 0:50)

***R**esearch in*
***A**stronomy and*
***A**strophysics*

Identifying symbiotic stars with machine learning

Yongle Jia[1], Sufen Guo[1*], Chunhua Zhu[1], Lin Li[1], Mei Ma[1] and Guoliang Lü[2,1*]

[1] School of Physical Science and Technology, Xinjiang University, Urumqi 830046, China; *guosufen@xju.edu.cn*

[2] Xinjiang Astronomical Observatory, Chinese Academy of Sciences, 150 Science 1-Street, Urumqi, Xinjiang 830011, China; *guolianglv@xao.ac.cn*

Received 2023 ****; accepted 2023 ****

Abstract Symbiotic stars are interacting binary systems, making them valuable for studying various astronomical phenomena, such as stellar evolution, mass transfer, and accretion processes. Despite recent progress in the discovery of symbiotic stars, a significant discrepancy between the observed population of symbiotic stars and the number predicted by theoretical models. To bridge this gap, this study utilized machine learning techniques to efficiently identify new symbiotic stars candidates. Three algorithms (XGBoost, LightGBM, and Decision Tree) were applied to a dataset of 198 confirmed symbiotic stars and the resulting model was then used to analyze data from the LAMOST survey, leading to the identification of 11,709 potential symbiotic stars candidates. Out of the these potential symbiotic stars candidates listed in the catalog, 15 have spectra available in the SDSS survey. Among these 15 candidates, two candidates, namely V* V603 Ori and V* GN Tau, have been confirmed as symbiotic stars. The remaining 11 candidates have been classified as accreting-only symbiotic star candidates. The other two candidates, one of which has been identified as a galaxy by both SDSS and LAMOST surveys, and the other identified as a quasar by SDSS survey and as a galaxy by LAMOST survey.

Key words: binaries: symbiotic — techniques: spectroscopic — binaries: spectroscopic — methods: data analysis

1 INTRODUCTION

Symbiotic stars are a unique type of binary system, characterized by prolonged interactions between the two stars (Lü et al. 2006, 2009, 2012; Han et al. 2020). These systems are composed of three components: a hot companion such as a white dwarf, neutron star, or main sequence star with an accretion disk; a cool companion, such as a red giant or asymptotic giant branch star; and an ionized nebula formed from material lost by the cool companion (Kenyon et al. 1991; Munari et al. 2021). The spectrum of symbiotic stars is composed of common features, including emission lines from the hot companion, absorption lines from the cool companion, and the nebula's material. In many symbiotic stars, the cool companion loses mass through stellar wind or Roche-lobe overflow, while the hot companion accretes enough material to produce the symbiotic phenomenon (Mikołajewska 2007; Akras et al. 2021; Iłkiewicz et al. 2022). Symbiotic stars provide an excellent sample for studying the loss of matter, acceleration mechanisms of stellar winds, and accretion of stellar winds in late-type giants (Chen et al. 2017; Saladino et al. 2019). They are also considered to be the precursors of Ia supernovae and important sources of soft and hard X-rays (Munari & Renzini 1992).

Since symbiotic stars are unique astrophysical laboratories. Allen (1984) provided a catalog of symbiotic stars, including 129 symbiotic stars and 15 symbiotic stars candidates. Kenyon (2009) summarized a catalog of symbiotic stars, including 133 symbiotic stars and 20 symbiotic stars candidates. Belczyński et al. (2000) provided a more detailed catalog of 188 symbiotic stars and 30 symbiotic stars candidates. Akras et al. (2019a) provided a catalog of 323 known symbiotic stars, 257 are Galactic and 66 extragalactic. Akras (2023) obtained 814 symbiotic star candidates by GALEX UV and 2MASS/AllWISE IR photometry, and identified two symbiotic stars after spectral analysis. Akras et al. (2019a) proposed a new subfamily of symbiotic stars for the first time: the S+IR type.

The symbiotic stars are classified into four categories: S, D, D', and S+IR (Akras et al. 2019a). Akras et al. (2019a) used a blackbody radiation model to fit the spectral energy map of symbiotic stars and found a number of S-type symbiotic stars with an S-type SED profile and an infrared excess in the mid-infrared regime. Akras et al. (2019a) named these symbiotic stars S+IR type. Most of the symbiotic stars are S-type, and the SED profiles of S-type symbiotic stars peak at 0.8 to 1.7 μm with an average value of 1.07 μm. Of course, a small percentage of S-type symbiotic stars with SED diagrams peaking around 0.7 or 1.8 μm. The temperature is between 3400 and 3800 K. Thirteen percent of the symbiotic stars are D-type, and the SED diagram for D-type symbiotic stars peaks at 2 to 4 μm with an average value of 2.85 μm. The number of D'-type symbiotic stars is relatively small, with only 10 known symbiotic stars with G/K spectral type. The SED diagram of D'-type symbiotic stars has two peaks between 2 and 10 μm. The SED diagram of S+IR-type symbiotic stars has a peak at 1.3 μm. The IR excess indicates the presence of a dusty shell around the symbiotic star and a lower temperature than that of the D-type. It is most likely that there is an accretion disk around the white dwarf (Akras et al. 2019a). Currently, there is a lack of in-depth research on the spectral characteristics of S+IR-type symbiotic stars. However, Gutierrez-Moreno & Moreno (1996); Pereira et al. (1998); Mürset & Schmid (1999); Pereira et al. (2005); Kogure & Leung (2007); Stoyanov et al. (2020) had conducted research on the spectral characteristics of S-type, D-type and D'-type symbiotic stars. The latest list of symbiotic stars was summarized by Merc et al. (2020), with 275 symbiotic stars and 119 symbiotic stars candidates in the galaxy. Although the number of observed symbiotic stars is increasing, the number of observed symbiotic stars differs significantly from what was predicted. Lü et al. (2006) predicted that the number of symbiotic stars with white dwarf(WD) accretors in the Galaxy may range from about 1200 to 15000, and the model birth rate of symbiotic stars in the Galaxy is from 0.035-0.131 yr^{-1}. It is predicted that the number of symbiotic stars in the galaxy should be from 3×10^3 to 4×10^5 (Allen 1984; Magrini et al. 2003; Lü et al. 2006). The number of known symbiotic stars is very different from the predicted number, so we need to discover more symbiotic stars to provide astronomers with data. Previously, symbiotic stars were found by spectroscopic analysis. With the introduction of various telescopes, astronomical data has grown exponentially, and the previous methods of processing data are unsuitable for the current requirements of processing large data. We need a new method to discover new symbiotic stars from the huge amount of astronomical observation data.

Machine learning has become a well-established tool in the field of astronomy. For instance, Gulati et al. (1998) and Singh et al. (1998) employed artificial neural networks to classify stellar spectra, while Bu et al. (2014) applied support vector machines to achieve a more detailed classification of stellar spectra. Guo et al. (2018) utilized support vector machines to establish Wide-field Infrared Survey Explorer (WISE) mid-infrared color criteria for the selection of quasar candidates, resulting in the compilation of a catalog of quasar candidates. Fu et al. (2021) employed the Extreme Gradient Boosting (XGBoost) algorithm for machine learning to Pan-STARRSI (PSI) and AllWISE photometry in order to classify quasars behind the Galactic plane (GPQ). This led to the development of a reliable GPQ candidate catalog. Akras et al. (2021) utilized machine learning methods to discover five new symbiotic stars and proposed a novel method for their identification, thus rendering the determination of symbiotic stars no longer dependent on spectroscopic analysis alone. However, the method employed by Akras only distinguished symbiotic stars from other objects containing H emission lines. It is likely that other types of objects exist that have not been considered. Therefore, it is imperative to develop a model that can be applied more broadly. In this study, we expand upon this idea and propose a method that can be used to quickly identify symbiotic stars among a large number of objects. We used the aggregated coordinates

of symbiotic stars that were cross-matched with AllWISE and Two Micron All Sky Survey (2MASS) data for machine learning training, and then applied the trained machine model to identify new symbiotic stars in Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST). Through the application of machine learning, we were able to identify a new set of symbiotic stars candidates.

The paper is divided into several sections. in Section 2, we provide a detailed description of the data sources utilized, including AllWISE, 2MASS, LAMOST, and Sloan Digital Sky Survey (SDSS), and the composition of the training data. In Section 3, we discuss the machine learning models selected and the training process. In Section 4, we present our prediction results and the spectral analysis of two newly-discovered symbiotic stars. Lastly, in Section 5 we summarize our results.

2 DATA

2.1 The AllWISE catalog

The WISE is a medium class explorer mission funded by NASA (Duval et al. 2004; Wright et al. 2010; Liu et al. 2008). WISE uses a 40 cm telescope to image the entire sky in four infrared bands W1, W2, W3, and W4 at 3.4, 4.6, 12, and 22 μm and has already produced over a million images and hundreds of millions of celestial bodies have been observed. The AllWISE catalog (Cutri et al. 2021) extends the results of the Wide-field Infrared Survey Mission (Cutri & et al. 2012). It combines data from the cryogenic and post-cryogenic periods to provide the most comprehensive mid-infrared overview currently available. The AllWISE Source Catalog contains accurate positions, motion measurements, photometry and ancillary information for 747,634,026 objects (Cutri et al. 2013). The AllWISE survey has yielded better photometric sensitivity, accuracy, and astrometric precision data than the WISE survey.

2.2 The 2MASS catalog

The 2MASS observes the sky in the near-infrared J (1.25 μm), H (1.65 μm), and Ks (2.16 μm) band separately, covering 99.998% of the sky observed from both the northern 2MASS facility at Mt. Hopkins, AZ, and the southern 2MASS facility at Cerro Tololo, Chile(Cutri et al. 2003;). The release data products include 4,121,439 Atlas Images in the three survey bands, and Catalogs containing positional and photometric information for 470,992,970 Point sources and 1,647,599 Extended sources(Kleinmann 1992; Kleinmann et al. 1994; Skrutskie et al. 2006).

2.3 The LAMOST catalog

The LAMOST is a new type of wide field of view and large aperture telescope that is a special quasi-meridian reflecting Schmidt telescope located in Xinglong Station of the National Astronomical Observatory, China (Cui et al. 2012; Zhao et al. 2012; Luo et al. 2015). LAMOST is an international leader in the field of wide-field optical spectroscopy and astronomy. It observes astronomical spectra in the northern sky. LAMOST began its first spectroscopic survey in September 2012 and has released its tenth data release (LAMOST DR 10 v0), containing more than 11 million spectra of 10 million stars, 242,569 galaxies, and 76,167 quasars.

2.4 The SDSS catalog

The SDSS (Gunn et al. 1998; Finlator et al. 2000; York et al. 2000) is a wide-field optical/infrared imaging and spectroscopy survey using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory. SDSS started in 1998 and has completed four phases: I, II, III, IV and V. Currently, the SDSS 18th data (DR18) has been released (Almeida et al. 2023). SDSS provides images, spectra, and scientific catalogs. To date, more than 6 million spectra have been observed by SDSS (Abdurro'uf et al. 2022).

2.5 Feature selection

The 2MASS J-H versus H-Ks color-color distribution diagrams have been widely used to investigate the near-infrared properties of symbiotic stars, and to differentiate between S-type, D-type, or other new candidates(Allen & Glass 1974; Rodríguez-Flores et al. 2014; Baella et al. 2013). Baella et al. (2016) discovered that the W3-W4 color index from the WISE survey can differentiate normal K giants from D-type symbiotic stars. Akras et al. (2019b) also utilized the magnitudes from both 2MASS and WISE surveys to distinguish symbiotic stars from other objects with strong H emission lines. Therefore, in our study, we used magnitudes from seven bands (W1, W2, W3, W4, J, H, and Ks) as the features for model training.

2.6 Data set composition

Currently, Merc et al. (2020) summarized the previously discovered and confirmed symbiotic stars into a new symbiotic star catalog. As of September 2022, this catalog contains 275 Galactic and 200 extragalactic symbiotic stars. This is by far the most comprehensive catalog of symbiotic stars. Since a lot of observational data about stars in the Milky Way, we only use the total of 275 Galactic symbiotic stars summarized in this catalog. We cross-matched the aggregated 275 symbiotic stars with AllWISE at a radius of 10 arcsec due to the angular resolution of the WISE survey, which is $6.1''$, $6.4''$, $6.5''$ & $12.0''$ at 3.4, 4.6, 12 & 22 μm (Wright et al. 2010). The cross-match radius between the target source and the matched source is shown in Figure 1. The sources in the catalog are cross-matched to their corresponding sources in AllWISE, and 82.5% (227) of them are cross-matched to sources in AllWISE with a radius less than 0.4 arcsec, only 3 sources were found to have a radius greater than 10 arcsec. We believe that these 3 sources do not have corresponding sources in AllWISE. We performed cross-matching operations using TOPCAT (Taylor 2005). The AllWISE catalog has been cross-matched with 2MASS with a matching radius of 3 arcsec (Cutri et al. 2013). Therefore, when cross-matching a catalog with the AllWISE catalog, the resulting match will include magnitude information from the 2MASS catalog. Two symbiotic stars were missing some magnitude information, thus leaving a total of 270 symbiotic stars available for training. We randomly divided the 270 symbiotic stars available for training into two parts. The first part consisted of 198 symbiotic stars used as positive samples to compose the data set with non-symbiotic stars. The second part consisted of 72 symbiotic stars used separately to test the trained models.

We used the LAMOST DR9 V1.0 spectroscopic dataset, which comprises a vast collection of 10,907,516 star spectra, 242,569 galaxy spectra, and 76,167 quasar spectra. To compose a training set, we employed a random selection strategy to choose one spectrum from every 500 star spectra, one spectrum from every 20 galaxy spectra, and one spectrum from every 3 quasar spectra. We cross-matched the selected sources with AllWISE using a matching radius of 1 arcsec to ensure accurate matches. We removed spectra without 2MASS magnitudes, resulting in a dataset of 12,348 star spectra, 9,014 galaxy spectra, and 3,797 quasar spectra. Finally, we composed the dataset by using these non-symbiotic stars as negative samples together with the 198 positive samples.

Due to the significant difference in the number of symbiotic stars and non-symbiotic stars in the training set, we randomly selected an additional 200 non-symbiotic stars from the non-symbiotic star samples in the training set as additional training set samples. Additionally, boxplots were used to depict the distribution of the training and test sets in different bands, as illustrated in Figure 2. The horizontal orange line inside the box in the boxplot represents the median of the sample data for the corresponding band. The upper and lower limits of the box represent the upper quartile and lower quartile of the sample data for the corresponding band, respectively. There is a line above and below the box, which represents the maximum and minimum values of the sample data in the corresponding band. Hollow dots indicate some outliers in the sample data in the corresponding band. As illustrated in Figure 2 , it can be observed that the two parts of the data exhibit a homogeneous distribution.

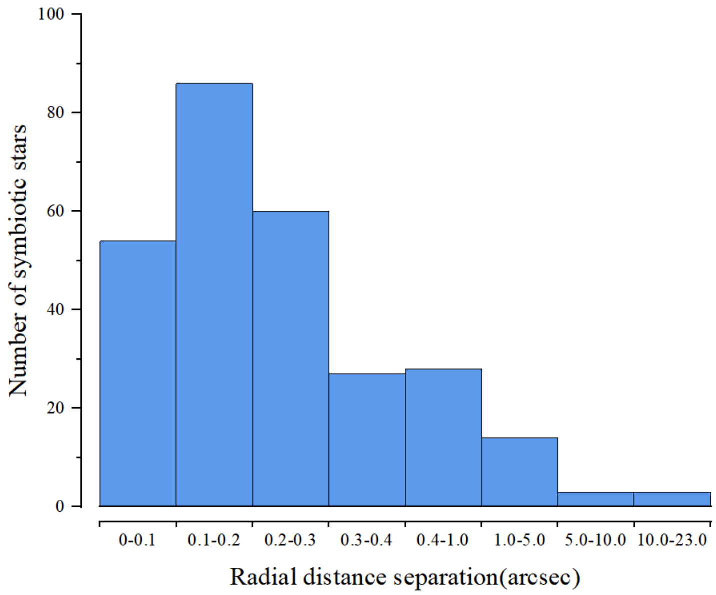

Fig. 1: The radial distance between the 275 symbiotic stars and the matched sources in AllWISE.

3 MODEL TRAINING

3.1 Model selection

In this work, we utilized three machine learning algorithms: the Decision Tree algorithm, the XGBoost algorithm, and the Light Gradient Boosting Machine (LightGBM) algorithm.

The Decision Tree algorithm is a method for approximating the value of a discrete function (Rokach & Maimon 2005; Barros et al. 2012), and is a commonly used classification technique. It begins by processing the data and generating a set of readable rules and decision trees through an induction algorithm. These rules are then used to analyze new data, effectively classifying it. This algorithm is a popular method for predictive modeling in fields such as statistics, data mining, and machine learning (Kotsiantis & S. 2013). The Decision Tree algorithm is a tree-like structure that can be used to solve classification and regression problems. The construction process of a decision tree is based on the training set. Firstly, a feature is chosen as the root node, and then the dataset is divided into several subsets. Each subset of data has the same feature value. Then, a feature is selected as a node for classification in each data subset in turn. This process is repeated until a complete decision tree is generated. The advantage of a decision tree is that it is easy to understand and interpret, and the classification rules can be clearly understood. However, decision trees also suffer from the problem of overfitting, which needs to be optimized through methods such as pruning (Barros et al. 2012). Vasconcellos et al. (2011) explored 13 Decision tree algorithms for star/galaxy classification using SDSS DR7 data and proposed a novel method that can accurately distinguish between stars and galaxies with high precision.

The XGBoost algorithm, proposed by Tianqi Chen at the University of Washington, is a powerful tool for classification and regression tasks (Chen & Guestrin 2016; Wang et al. 2020). It is composed of multiple Classification And Regression Tree decision trees, each of which learns the residual of the target value and the sum of all previous tree predictions. The final prediction is made by combining the results of all the trees (Friedman et al. 2000). Though each weak classifier may not have a high global prediction accuracy, they can still have a very high prediction accuracy for specific aspects of the data(Chen & Guestrin 2016). By combining many classifiers with high local prediction accuracy, the XGBoost algorithm can achieve the effect of a strong classifier with high global prediction accuracy. It has gained popularity in data modeling competitions due to its excellent computing efficiency and

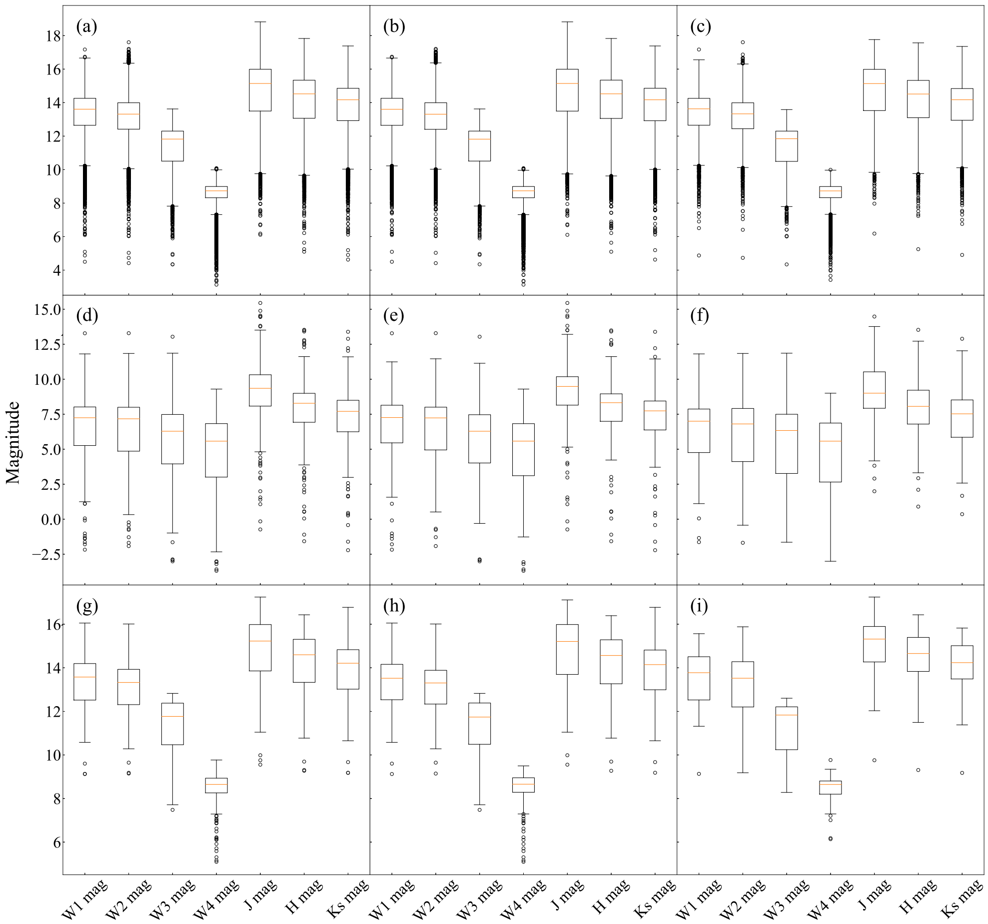

Fig. 2: The sequence of boxplots displays the magnitude distribution of different sets of stars. Specifically, Boxplots (a) exhibit the magnitude distribution of 25,159 non-symbiotic stars, with Boxplots (b) and (c) presenting the magnitude distribution of the subset assigned to the training and testing sets, respectively. Similarly, Boxplots (d) show the magnitude distribution of 270 symbiotic stars, with Boxplots (e) and (f) displaying the magnitude distribution of the subset assigned to the training and validation sets, respectively. Lastly, Boxplots (g) illustrate the magnitude distribution of 200 randomly selected non-symbiotic stars, with Boxplots (h) and (i) depicting the magnitude distribution of the subset assigned to the training and testing sets, respectively. The horizontal orange line inside the box in the boxplot represents the median of the sample data for the corresponding band. The upper and lower limits of the box represent the upper quartile and lower quartile of the sample data for the corresponding band, respectively. There is a line above and below the box, which represents the maximum and minimum values of the sample data in the corresponding band. Hollow dots indicate some outliers in the sample data in the corresponding band.

prediction accuracy. It is widely used in various fields such as identifying stars, galaxies, and quasars from Beijing-Arizona Sky Survey (BASS) DR3 with accuracy more than 90% and classifying stars and galaxies using different models from SDSS DR7 (Li et al. 2022; Li et al. 2019).

The LightGBM algorithm is a scalable machine learning system developed by Microsoft in 2017. It is an open-source project led by Guolin Ke, a winner of the first Alibaba Big Data Competition in 2014. The algorithm is based on the Gradient Boosting Decision Tree and is designed to reduce mem-

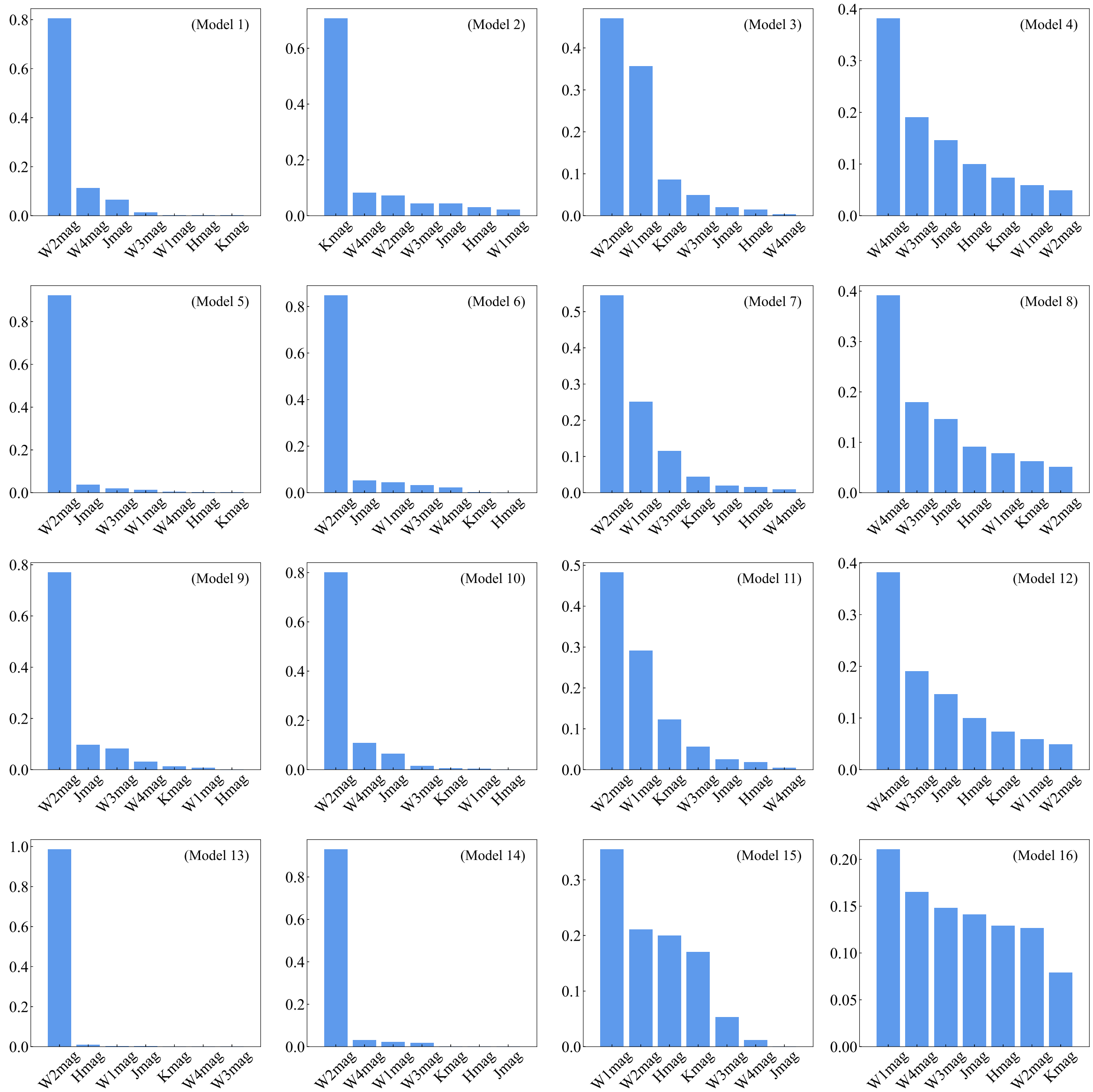

Fig. 3: Distribution chart of feature importance for 16 best machine learning models.

ory and computational requirements while also minimizing communication costs when used in parallel with multiple machines (Ke et al. 2017). LightGBM can automatically identify effective data features and is considered an improved version of the XGBoost algorithm, providing faster training speed and lower memory consumption while maintaining a similar level of accuracy (Wang et al. 2020). Due to its efficient performance on large-scale datasets, LightGBM has become a popular tool in data competitions. Malik et al. 2022 used LightGBM to identify planets in simpler data and Transiting Exoplanet Survey Satellite (TESS) data, with very good results.

3.2 Data Balancing Algorithms

During machine learning training, it has been observed that when the number of negative samples exceeds that of positive samples, the model focuses more on the characteristics of negative samples. Therefore, it is necessary to use some methods to reduce the issues caused by the serious imbalance between the number of positive and negative samples. To address this issue, we employed the Synthetic Minority Oversampling Technique (SMOTE) algorithm (Castellanos et al. 2018) and the Edited Nearest Neighbours (ENN) algorithm (Wilson 1972) to balance the dataset.

SMOTE is an algorithm used to address imbalanced classification problems. It generates synthetic samples of the minority class by interpolating between existing samples. The steps are: select a sample, identify its k nearest neighbors, randomly select one neighbor, generate a new synthetic sample, and repeat until the desired number of samples is reached. The new samples are combined with the original minority class samples to create a balanced dataset (Chawla et al. 2002).

ENN is an undersampling technique commonly employed for addressing class imbalance by reducing the majority class to match the minority class (Tang & He 2015). In order to determine whether the majority class of an observation's k nearest neighbors is the same as the class of the observation itself, the ENN technique identifies the k nearest neighbors for each observation. If the majority class of an observation's k nearest neighbors differs from the class of the observation, both the observation and its k nearest neighbor are removed from the dataset (Kim 2021).

SMOTE and ENN algorithms were applied to the training data, resulting in the construction of 12 machine learning models. To compare with the generated 12 models and accurately identify symbiotic stars, we constructed a new training set with a nearly 1:1 ratio of positive and negative samples. We randomly selected 200 non-symbiotic stars from the training set and combined them with our 198 symbiotic stars to create the new training set. We applied the Decision Tree, XGBoost, and LightGBM algorithms to the new training set, resulting in the construction of a total of 16 machine learning models throughout our research process. Table 1 shows the algorithms and constructions used for the 16 models.

3.3 Results

We randomly partitioned the dataset into a training set consisting of 80% of the data set and a test set consisting of 20% of the data set, with the aim of ensuring a randomized partition. The training set was utilized for training the machine learning models, while the test set was employed for evaluating the performance of the machine learning models. We trained the models with Scikit-learn. Scikit-learn is a Python-based open-source machine learning library that offers an extensive collection of tools and algorithms for data analysis, including classification, regression, clustering, dimensionality reduction, and model selection (Pedregosa et al. 2011;Buitinck et al. 2013). During the training process, we employ 5-fold cross-validation and GridSearchCV to optimize the model parameters. Cross-validation is a statistical analysis method commonly used to evaluate prediction performance and maximize data utilization (Yadav & Shukla 2016). GridSearchCV is a traditional approach for hyperparameter optimization in machine learning (Müller & Guido 2016).

To evaluate the performance of our machine learning models, we used four key evaluation metrics: Precision, Recall, F1-Score, and AUC. These metrics are defined as follows:

1. Precision: The proportion of true positive samples among the predicted positive samples. It is calculated as TP / (TP + FP).
2. Recall: The proportion of true positive samples that are correctly predicted as positive. It is calculated as TP / (TP + FN).
3. F1-Score: The harmonic mean of precision and recall. It is calculated as 2 * (precision * recall) / (precision + recall).
4. AUC: The area under the ROC curve, which can be used to measure the performance of a classification model. The AUC value ranges from 0 to 1, with a higher value indicating better model performance.

Table 1: Summarizing the 16 best performing machine learning models

	Decision Tree_gini	Decision Tree_entropy	XGBoost	LightGBM
Raw dataset	Model 1	Model 2	Model 3	Model 4
SMOTE	Model 5	Model 6	Model 7	Model 8
ENN	Model 9	Model 10	Model 11	Model 12
200 non-symbiotic stars	Model 13	Model 14	Model 15	Model 16

Table 2: The performance of the 16 best models in predicting symbiotic stars on the test set, presented as Precision, Recall, F1-Score, and the AUC scores.

	model 1	model 2	model 3	model 4	model 5	model 6	model 7	model 8
Precision	0.94	0.94	0.92	0.94	0.96	0.92	0.91	0.94
Recall	0.88	0.88	0.88	0.92	0.92	0.94	0.98	0.90
F1-Score	0.91	0.91	0.90	0.93	0.94	0.93	0.94	0.92
AUC	0.94	0.94	0.94	0.96	0.96	0.97	0.99	0.95
	model 9	model 10	model 11	model 12	model 13	model 14	model 15	model 16
Precision	0.94	0.96	0.94	0.94	0.97	0.97	0.97	1.00
Recall	0.96	0.87	0.92	0.92	0.97	0.97	0.97	0.97
F1-Score	0.95	0.91	0.93	0.93	0.97	0.97	0.97	0.99
AUC	0.98	0.93	0.96	0.96	0.97	0.97	0.97	0.99

Table 3: The results presented in the table are the predictions made by the 16 trained models for 72 known symbiotic stars (SySts)

	model 1	model 2	model 3	model 4	model 5	model 6	model 7	model 8
$\hat{y}$ = SySts	60	59	63	62	59	62	66	63
	model 9	model 10	model 11	model 12	model 13	model 14	model 15	model 16
$\hat{y}$ = SySts	62	53	64	62	70	70	70	70

Here, TP represents the number of samples predicted by the model as symbiotic stars and are actually symbiotic stars. FP represents the number of samples predicted by the model as symbiotic stars but are actually non-symbiotic stars. FN represents the number of samples predicted by the model as non-symbiotic stars but are actually symbiotic stars. The primary evaluation metric for our trained model in this study was Precision, as our goal was to accurately identify a maximum number of symbiotic stars.

The performance of 16 models on the test set is summarized in Table 2. For a comprehensive understanding of the parameters of these models, please refer to Appendix A. Additionally, Figure 3 presents the feature importance of the trained models. From Figure 3, it can be observed that the feature importance of W2 is very high in the models that used decision trees (Models 1, 2, 5, 6, 9, 10, 13, and 14). The feature importance of other magnitudes is very low, indicating that decision tree models, although easy to construct and structurally simple, may not be stable enough. In the models that utilized XGBoost (Models 3, 7, 11, and 15), W2 and W1 have relatively high feature importance. For the models that employed LightGBM (Models 4, 8, 12, and 16), the feature importance of W4 and W3 is relatively high. The performance of the trained models was evaluated using a separate set of 72 symbiotic stars. Table 3 provides a clear view of the predictions made by each model for these 72 symbiotic stars.

4 IDENTIFYING SYMBIOTIC STARS

The most reliable method for determining symbiotic stars is through spectroscopic analysis. The earliest criteria for identifying symbiotic stars were proposed by Merrill & Humason (1932), stating that a symbiotic star is a binary system with a combined spectrum. As the study of symbiotic stars progressed, more refined criteria for identifying symbiotic stars were proposed, with the most current and widely

accepted criteria being those proposed by Belczyński et al. (2000). However, it is important to note that these criteria are only applicable to burning symbiotic stars, and accreting-only symbiotic stars may not show these spectral features unless an outburst occurs (Merc et al. 2021). In this study, we used the criteria proposed by Belczyński et al. (2000) to classify symbiotic stars. These criteria are as follows:

1. resence spectral features of late-type giants such as TiO, $\mathrm{H_2O}$, CO, CN and VO bands as well as CaI, CaII, FeI and NaI absorption lines.
2. the detection of some typical emission lines, for instance HI and HeI. emission lines of ions with an ionization potential of at least 35 eV (e.g. [OIII], [FeVII]$\lambda\lambda$5727,6087, [HeII]λ4686).
3. the presence of strong emission lines of $\mathrm{OVI}\lambda\lambda$6830,7088 (Mikolajewska et al. 1997; Belczyński et al. 2000; Akras et al. 2019a).

Akras et al. (2019b) proposed a new method for determining symbiotic stars using a mid-infrared criterion. This criterion was used to successfully discover five previously unknown symbiotic stars in the galaxy (Akras et al. 2021).

We aimed to classify symbiotic stars among a sample of 11,226,252 sources from LAMOST DR9 v1.0 using 16 machine learning models. We cross-matched the sources with the AllWISE and 2MASS catalogs at a radius of 6 arcsec, and obtained magnitude data for a total of 10,849,157 sources. The models were then utilized to identify symbiotic stars. In order to indentifying more reliable symbiotic star prediction, we employed a methodology that integrates the results of multiple models to verify the classification of a star as a symbiotic star. This ensures that the star is identified as a symbiotic star by all of the models used in our analysis.

A total of 11,709 sources were jointly identify as symbiotic stars by the 16 models. These sources were then cross-matched with SDSS DR17 at a radius of 6 arcsec, and 15 of them were found to have spectral information. Among these 15 candidates, one was classified as a galaxy by both surveys, while another was identified as a quasar by SDSS and a galaxy by LAMOST, and 13 were found to have spectra similar to symbiotic stars. We have created a catalog of the relevant information for these sources and made it available on the website[1]. The spectra of 2 of the extent 13 sources are more in line with the criteria of Belczyński et al. (2000) for determining the spectra of symbiotic stars. The spectra of the extent 11 sources are similar to the accreting-only symbiotic star V562Lyr.

Here we discussed the spectra of the two sources that have been newly predicted as symbiotic stars.

4.1 V* V603 Ori

V* V603 Ori (= 2MASS J05393983-0233160; RA2000 = 05 39 39.8292920832, DEC2000 = -02 33 16.040160936; see also Table 4) displays more prominent $\mathrm{H}\alpha$, HeI lines, which increases the likelihood that it is a symbiotic star. Additionally, the presence of more distinct HeII, $\mathrm{H}\beta$, HeI λ4930, OIII λ5007, and also $\mathrm{H}\gamma$(Figure 4) lines further supports this possibility. However, V* V603 Ori differs from the newly discovered symbiotic star DR2J141301.4-6533201.1 in Akras et al. (2021), as the latter lacks the OIII λ4363 line, yet is still considered a symbiotic star. V* V603 Ori also exhibits some spectral characteristics of red giants, such as a molecular band containing TiO and absorption Na lines, as well as emission lines likeNeIII. According to the magnitude information of AllWISE and 2MASS of V* V603 Ori, it is classified as an S+IR-type symbiotic star. However, the TESS survey has not captured any obvious light variation in V* V603 Ori, thus it cannot be confirmed as a symbiotic star through light variation analysis. Fortunately, the renormalised unit weight error (RUWE) parameter in GaiaDR3 is an essential indicator for determining whether a star is binary. According to Hambly et al. (2021), when the RUWE value is 1.4, it is likely that the star is not single. In the case of V* V603 Ori, the RUWE value is 1.4. Iłkiewicz & Mikołajewska (2017) explored various other emission lines including [N II], [O III], [Ne III] and HeI lines, in order to distinguish planetary nebulae from symbiotic stars and proposed criteria for symbiotic stars based on emission line ratios. for example, $\log(\text{[O III] }5006/\text{[N II] }5755) < 2.6$ and $\log(\text{[Ne III] }3869/\text{[O III]}4363) < 0.45$. V* V603 Ori meets

[1] https://doi.org/10.12149/101183

Table 4: Basic properties of V* V603 Ori and V* GN Tau. Data are from Gaia DR3 (Gaia Collaboration et al. 2023), 2MASS (Skrutskie et al. 2006), and AllWISE (Wright et al. 2010).

Soucre	RA (J2000)	DEC (J2000)	J (mag)	H (mag)	Ks (mag)	$W1$ (mag)	$W2$ (mag)	$W3$ (mag)	$W4$ (mag)
V* V603 Ori	84.915955	-02.554456	12.22 ± 0.03	10.96 ± 0.02	10.07 ± 0.02	8.96 ± 0.02	8.33 ± 0.03	6.76 ± 0.02	5.05 ± 0.04
V* GN Tau	69.837176	+25.750563	10.20 ± 0.03	8.89 ± 0.03	8.06 ± 0.03	7.16 ± 0.04	6.53 ± 0.02	4.91 ± 0.01	3.05 ± 0.02

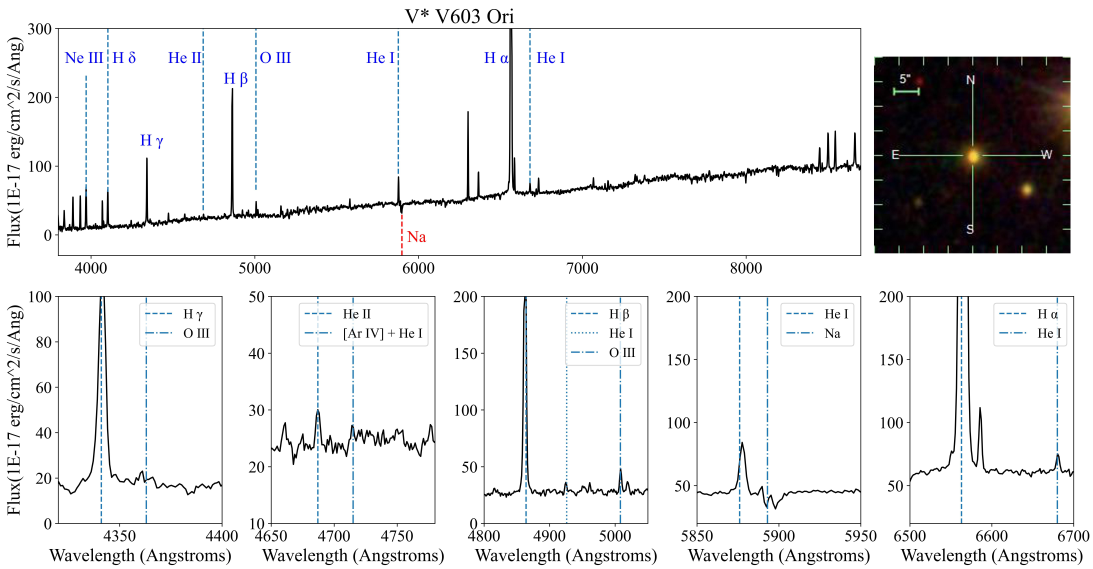

Fig. 4: Low resolution spectra of V* V603 Ori from SDSS. Top left panel shows the V* V603 Ori observed spectra in SDSS. The top right panel displays the images of the V* V603 Ori in SDSS. South is down, west to the right. Bottom panels zoom in the Hγ and OIII 4363Å lines, HeII 4686Å line, Hβ and OIII 5007Å lines, Hα and HeI emission lines.

this criterion. V* V603 Ori has been classified as an M0 spectral type star by LAMOST. The symbiotic stars were cross-matched with the Gaia catalog within 5 arcsec , and corresponding diagnostic colour-colour diagrams were plotted for The Gaia G_{BP}-G versus G_{BP}-G_{RP}. In the figure5, V* V603 Ori was highlighted in red. Based on the above information, we classify V* V603 Ori as a newly discovered symbiotic star.

4.2 V* GN Tau

The source J043920.90+254502.1 in LAMOST cross-matched with AllWISE is V* GN Tau(see also Table 4), which is predicted to be a symbiotic star by our model. The SDSS spectrogram of V* GN Tau displays prominent Hα and HeI lines (Figure 6), suggesting that it is a strong candidate for being a symbiotic star. Other noticeable lines in the spectrum include HeII ,Hβ, λ4930, OIII λ5007, and also Hγ and NaI absorption, as well as a molecular band of TiO. Although there is a lack of OIII λ4363 lines, it does not necessarily exclude the possibility of V* GN Tau being a symbiotic star. Other known symbiotic stars such as AS 201, V3811 Sgr, and V407 Cyg also display a similar absence of certain spectral lines (refer to Table A.7 in Kenyon (2009)). The spectrum of V* GN Tau is similar to that of DR2J175346.2-284826.16, which was previously identified as a symbiotic candidate by Akras et al. (2021).

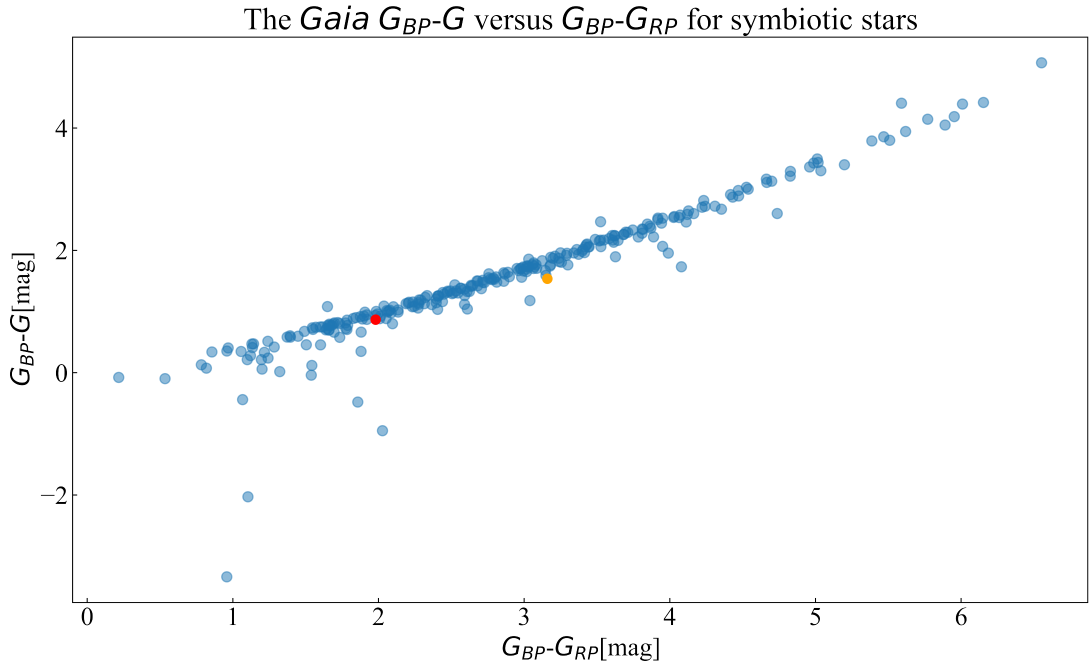

Fig. 5: The *Gaia* G_{BP}-G versus G_{BP}-G_{RP} for symbiotic stars. The sources were identified through cross-matching 275 symbiotic stars with the Gaia catalog within a radius of 5 arcsec and are represented in blue. V* V603 Ori and V* GN Tau are shown in red and orange, respectively.

The magnitude information of V* GN Tau in AllWISE and 2MASS supports its classification as an S+IR-type symbiotic star according to Akras' mid-infrared color standards. However, the lack of clear light variation in V* GN Tau as captured by the TESS survey hinders us from fully confirming its status as a symbiotic star through photometric analysis. The RUWE parameter value of V* GN Tau is not available in GaiaDR3, thus it is not possible to determine whether V* GN Tau is a binary star. V* GN Tau also meets Iłkiewicz & Mikołajewska (2017) criterion. V* GN Tau has been classified as an M4/3 spectral type star by LAMOST. In the figure5, V* GN Tau was highlighted in orange. But V* GN Tau is still identified as a symbiotic star based on its spectral information.

5 DISCUSSION AND CONCLUSIONS

In this study, the AllWISE and 2MASS magnitude data of known symbiotic stars were used to identify potential symbiotic stars through a machine learning approach. The constructed model processed 10,849,157 AllWISE and 2MASS catalogs from LAMOST DR9 sources and identified 11,709 symbiotic stars.

We analyzed the SDSS spectra of these symbiotic star candidates and discovered two new symbiotic stars, namely V* V603 Ori and V* GN Tau. The spectra of these two symbiotic stars were found to be consistent with the spectral classification criteria of Belczyński et al. (2000) and the mid-infrared color classification criteria of Akras et al. (2019b).

Actually, the criteria for identifying symbiotic stars based on their spectra are typically used for active, or "burning" symbiotic stars. However, some symbiotic stars can remain hidden for a long period of time, referred to as accreting-only symbiotic stars (Pujol et al. 2023). These stars do not exhibit the spectroscopic characteristics of active symbiotic stars and do not have distinct emission lines in the optical band, but instead, are bright in the ultraviolet spectrum (Luna et al. 2013; Mukai et al. 2016; Munari et al. 2021; Merc et al. 2021). An example of this is V1988 Sgr, whose spectrum was analyzed

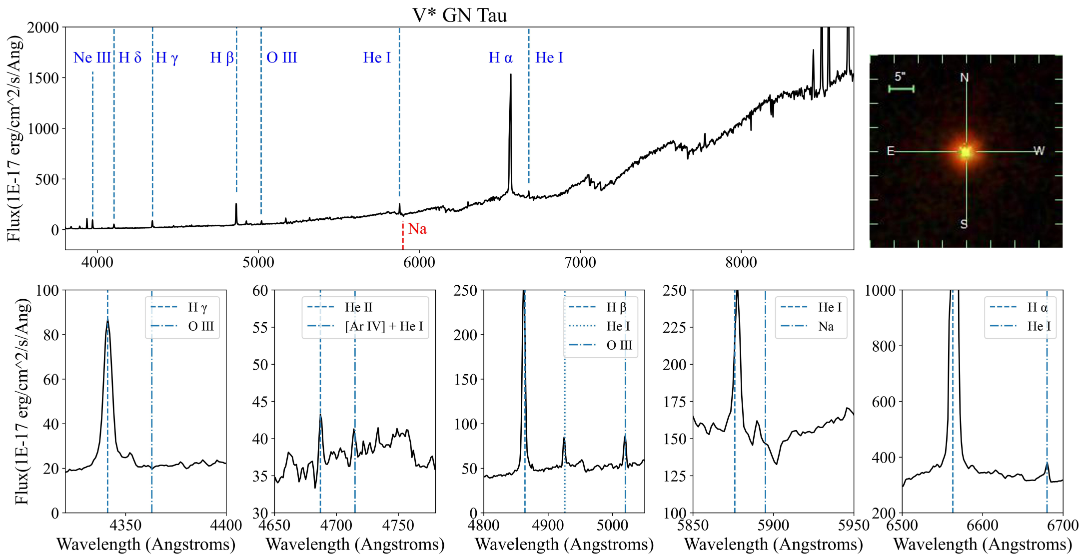

Fig. 6: Low resolution spectra of V* GN Tau Ori from SDSS. Top left panel shows the V* GN Tau observed spectra in SDSS. The top right panel displays the images of the V* GN Tau in SDSS. South is down, west to the right. Bottom panels zoom in the Hγ and OIII 4363Å lines, HeII 4686Å line, Hβ and OIII 5007Å lines, Hα and HeI emission lines.

by Merc et al. (2021) and found to not match the spectral features of active symbiotic stars. Despite this, Merc et al. (2021) concluded that V1988 Sgr could not be ruled out as an accreting-only symbiotic star. Another example is V562 Lyr, which was identified as a symbiotic star despite the absence of prominent H and He lines. This weakening or disappearance of optical emission lines in the spectra of symbiotic stars has been found to be due to a decrease in the accretion rate by Pujol et al. (2023).

Among the 11 remaining stars in our candidates catalog, we observed similar spectra, tentatively classifying them as accreting-only symbiotic star candidates. The next step in our research will be to observe these additional 11 candidates and gather more evidence to make a definitive classification. Further observations and data are necessary to confirm this classification.

Acknowledgements This work received the generous support of the Natural Science Foundation of Xinjiang No.2021D01C075, the National Natural Science Foundation of China, project Nos. 12163005, 12003025, U2031204 and 11863005, the science research grants from the China Manned Space Project with NO. CMS-CSST-2021-A10, the Scientific Research Program of the Higher Education Institution of Xinjiang (No. XJEDU2022P003). Data resources are supported by China National Astronomical Data Center (NADC) and Chinese Virtual Observatory (China-VO). This work is supported by Astronomical Big Data Joint Research Center, co-founded by National Astronomical Observatories, Chinese Academy of Sciences and Alibaba Cloud. This work made use of Astropy:[2] a community-developed core Python package and an ecosystem of tools and resources for astronomy (Astropy Collaboration et al. 2013; Astropy Collaboration et al. 2018; Astropy Collaboration et al. 2022). This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, and NEOWISE, which is a project of the Jet Propulsion Laboratory/California Institute of Technology. WISE and NEOWISE are funded by the National Aeronautics and Space Administration. This publication makes use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of

[2] http://www.astropy.org

Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. Guoshoujing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss4.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics — Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

References

Abdurro’uf, Accetta, K., Aerts, C., et al. 2022, ApJS, 259, 35 3
Akras, S. 2023, MNRAS, 519, 6044 2
Akras, S., Gonçalves, D. R., Alvarez-Candal, A., & Pereira, C. B. 2021, MNRAS, 502, 2513 1, 2, 10, 11
Akras, S., Guzman-Ramirez, L., Leal-Ferreira, M. L., & Ramos-Larios, G. 2019a, ApJS, 240, 21 2, 10
Akras, S., Leal-Ferreira, M. L., Guzman-Ramirez, L., & Ramos-Larios, G. 2019b, MNRAS, 483, 5077 4, 10, 12
Allen, D. A. 1984, PASA, 5, 369 2
Allen, D. A., & Glass, I. S. 1974, MNRAS, 167, 337 4
Almeida, A., Anderson, S. F., Argudo-Fernández, M., et al. 2023, arXiv e-prints, arXiv:2301.07688 3
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33 13
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123 13
Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167 13
Baella, N. O., Pereira, C. B., & Miranda, L. F. 2013, AJ, 146, 115 4
Baella, N. O., Pereira, C. B., Miranda, L. F., & Alvarez-Candal, A. 2016, AJ, 151, 100 4
Barros, R., Basgalupp, M., de Carvalho, A., & Freitas, A. 2012, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 42, 291 5
Belczyński, K., Mikołajewska, J., Munari, U., Ivison, R. J., & Friedjung, M. 2000, A&AS, 146, 407 2, 10, 12
Bu, Y., Chen, F., & Pan, J. 2014, New Astron., 28, 35 2
Buitinck, L., Louppe, G., Blondel, M., et al. 2013, in ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 108 8
Castellanos, F. J., Valero-Mas, J. J., Calvo-Zaragoza, J., & Rico-Juan, J. R. 2018, Pattern Recognition Letters, 103, 32 8
Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. 2002, J. Artif. Intell. Res. (JAIR), 16, 321 8
Chen, T., & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16 (New York, NY, USA: Association for Computing

Machinery), 785–794 5
Chen, Z., Frank, A., Blackman, E. G., Nordhaus, J., & Carroll-Nellenback, J. 2017, MNRAS, 468, 4465 1
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, Research in Astronomy and Astrophysics, 12, 1197 3
Cutri, R. M., & et al. 2012, VizieR Online Data Catalog, II/311 3
Cutri, R. M., Skrutskie, M. F., van Dyk, S., et al. 2003, 2MASS All Sky Catalog of point sources. 3
Cutri, R. M., Wright, E. L., Conrow, T., et al. 2013, Explanatory Supplement to the AllWISE Data Release Products, Explanatory Supplement to the AllWISE Data Release Products, by R. M. Cutri et al. 3, 4
Cutri, R. M., Wright, E. L., Conrow, T., et al. 2021, VizieR Online Data Catalog, II/328 3
Duval, V. G., Irace, W. R., Mainzer, A. K., & Wright, E. L. 2004, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 5487, Optical, Infrared, and Millimeter Space Telescopes, ed. J. C. Mather, 101 3
Finlator, K., Ivezić, Ž., Fan, X., et al. 2000, AJ, 120, 2615 3
Friedman, J., Hastie, T., & Tibshirani, R. 2000, The Annals of Statistics, 28, 337 5
Fu, Y., Wu, X.-B., Yang, Q., et al. 2021, ApJS, 254, 6 2
Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2023, A&A, 674, A1 11
Gulati, R. K., Gupta, R., & Singh, H. P. 1998, in ESA Special Publication, Vol. 413, Ultraviolet Astrophysics Beyond the IUE Final Archive, ed. W. Wamsteker, R. Gonzalez Riestra, & B. Harris, 711 2
Gunn, J. E., Carr, M., Rockosi, C., et al. 1998, AJ, 116, 3040 3
Guo, S., Qi, Z., Liao, S., et al. 2018, A&A, 618, A144 2
Gutierrez-Moreno, A., & Moreno, H. 1996, PASP, 108, 972 2
Hambly, N., Arenou, F., Babusiaux, C., et al. 2021, Gaia EDR3 documentation Chapter 13: Datamodel description, Gaia EDR3 documentation, European Space Agency; Gaia Data Processing and Analysis Consortium. Online at https://gea.esac.esa.int/archive/documentation/GEDR3/index.html, id. 13 10
Han, Z.-W., Ge, H.-W., Chen, X.-F., & Chen, H.-L. 2020, Research in Astronomy and Astrophysics, 20, 161 1
Iłkiewicz, K., & Mikołajewska, J. 2017, A&A, 606, A110 10, 12
Iłkiewicz, K., Mikołajewska, J., Scaringi, S., et al. 2022, MNRAS, 510, 2707 1
Ke, G., Meng, Q., Finley, T., et al. 2017, in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17 (Red Hook, NY, USA: Curran Associates Inc.), 3149–3157 7
Kenyon, S. J. 2009, The Symbiotic Stars 2, 11
Kenyon, S. J., Oliversen, N. A., Mikolajewska, J., et al. 1991, AJ, 101, 637 1
Kim, K. 2021, IEEE Access, 9, 143250 8
Kleinmann, S. G. 1992, in Astronomical Society of the Pacific Conference Series, Vol. 103, Robotic Telescopes in the 1990s, ed. A. V. Filippenko, 203 3
Kleinmann, S. G., Lysaght, M. G., Pughe, W. L., et al. 1994, Ap&SS, 217, 11 3
Kogure, T., & Leung, K.-C. 2007, The Astrophysics of Emission-Line Stars, Vol. 342 2
Kotsiantis, & S., B. 2013, Artificial Intelligence Review, 39, 261 5
Li, C., Zhang, W.-h., & Lin, J.-m. 2019, Chinese Astronomy and Astrophysics, 43, 539 6
Li, C., Zhang, Y., Cui, C., et al. 2022, MNRAS, 509, 2289 6
Liu, F., Cutri, R., Greanias, G., et al. 2008, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7017, Modeling, Systems Engineering, and Project Management for Astronomy III, ed. G. Z. Angeli & M. J. Cullum, 70170M 3
Lü, G. L., Zhu, C. H., Postnov, K. A., et al. 2012, MNRAS, 424, 2265 1
Lü, G., Yungelson, L., & Han, Z. 2006, MNRAS, 372, 1389 1, 2
Lü, G., Zhu, C., Wang, Z., & Wang, N. 2009, MNRAS, 396, 1086 1
Luna, G. J. M., Sokoloski, J. L., Mukai, K., & Nelson, T. 2013, A&A, 559, A6 12
Luo, A. L., Zhao, Y.-H., Zhao, G., et al. 2015, Research in Astronomy and Astrophysics, 15, 1095 3

Magrini, L., Corradi, R. L. M., & Munari, U. 2003, in Astronomical Society of the Pacific Conference Series, Vol. 303, Symbiotic Stars Probing Stellar Evolution, ed. R. L. M. Corradi, J. Mikolajewska, & T. J. Mahoney, 539 2
Malik, A., Moster, B. P., & Obermeier, C. 2022, MNRAS, 513, 5505 7
Merc, J., Gális, R., & Wolf, M. 2020, Contributions of the Astronomical Observatory Skalnate Pleso, 50, 426 2, 4
Merc, J., Gális, R., Wolf, M., et al. 2021, MNRAS, 506, 4151 10, 12, 13
Merrill, P. W., & Humason, M. L. 1932, PASP, 44, 56 9
Mikołajewska, J. 2007, Baltic Astronomy, 16, 1 1
Mikolajewska, J., Acker, A., & Stenholm, B. 1997, A&A, 327, 191 10
Mukai, K., Luna, G. J. M., Cusumano, G., et al. 2016, MNRAS, 461, L1 12
Müller, A. C., & Guido, S. 2016, Introduction to machine learning with Python: a guide for data scientists (" O'Reilly Media, Inc.") 8
Munari, U., & Renzini, A. 1992, ApJ, 397, L87 1
Munari, U., Traven, G., Masetti, N., et al. 2021, MNRAS, 505, 6121 1, 12
Mürset, U., & Schmid, H. M. 1999, A&AS, 137, 473 2
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825 8
Pereira, C. B., Landaberry, S. J. C., & Junqueira, S. 1998, A&A, 333, 658 2
Pereira, C. B., Smith, V. V., & Cunha, K. 2005, A&A, 429, 993 2
Pujol, A., Luna, G. J. M., Mukai, K., et al. 2023, A&A, 670, A32 12, 13
Rodríguez-Flores, E. R., Corradi, R. L. M., Mampaso, A., et al. 2014, A&A, 567, A49 4
Rokach, L., & Maimon, O. 2005, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 35, 476 5
Saladino, M. I., Pols, O. R., & Abate, C. 2019, A&A, 626, A68 1
Singh, H. P., Gupta, R., & Gulati, R. K. 1998, in Astronomical Society of the Pacific Conference Series, Vol. 138, 1997 Pacific Rim Conference on Stellar Astrophysics, ed. K. L. Chan, K. S. Cheng, & H. P. Singh, 309 2
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163 3, 11
Stoyanov, K. A., Iłkiewicz, K., Luna, G. J. M., et al. 2020, MNRAS, 495, 1461 2
Tang, B., & He, H. 2015, IEEE Computational Intelligence Magazine, 10, 52 8
Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29 4
Vasconcellos, E. C., de Carvalho, R. R., Gal, R. R., et al. 2011, AJ, 141, 189 5
Wang, M., Fu, W., He, X., Hao, S., & Wu, X. 2020, IEEE Transactions on Knowledge and Data Engineering, 34, 2574 5, 7
Wilson, D. L. 1972, IEEE Transactions on Systems, Man, and Cybernetics, SMC-2, 408 8
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868 3, 4, 11
Yadav, S., & Shukla, S. 2016, in 2016 IEEE 6th International Conference on Advanced Computing (IACC), 78 8
York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, AJ, 120, 1579 3
Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, Research in Astronomy and Astrophysics, 12, 723 3

Appendix A: ALGORITHM AND PARAMETER DETAILS OF THE 16 MACHINE LEARNING MODELS

Here is a detailed description of the parameters of these 16 models:

Models 1 through 4 were trained without using the SMOTE and ENN algorithms for data balancing. The training set consisted of 146 symbiotic stars and 20,139 non-symbiotic stars. Model 1 was a Decision Tree model that used the Gini index as the splitting criterion. Model 2 was a Decision Tree

model that used entropy as the splitting criterion. Model 3 was an XGBoost model, and Model 4 was a LightGBM model.

Models 5 through 8 were trained on a balanced dataset using the SMOTE algorithm. We conducted multiple rounds of parameter selection and determined the optimal parameter values for SMOTE algorithm using grid search, which are k_neighbors = 1 and sampling_strategy = 'minority'. After applying SMOTE, the training set consisted of 20,139 symbiotic stars and 20,139 non-symbiotic stars. Model 5 was a DecisionTree model trained on the SMOTE-balanced data using the Gini index as the splitting criterion, similar to Model 1. Model 6 was a DecisionTree model trained on the SMOTE-balanced data using the entropy criterion as the splitting criterion, similar to Model 2. Model 7 was an XGBoost model trained on the SMOTE-balanced data. Model 8 was a LightGBM model trained on the SMOTE-balanced data.

Models 9 through 12 were trained using the ENN algorithm to balance the data. We conducted multiple rounds of parameter selection and determined the optimal parameter values for ENN algorithm using grid search, which are kind_sel = 'mode', n_neighbors = 1 and sampling_strategy = 'majority'. After the application of ENN, the training set consisted of 20,120 symbiotic stars and 146 non-symbiotic stars. Model 9 is a DecisionTree model trained on the ENN-balanced data using the Gini index as the splitting criterion, similar to Model 1. Model 10 is a DecisionTree model trained on the ENN-balanced data using the entropy criterion as the splitting criterion, similar to Model 2. Model 11 is an XGBoost model trained on the ENN-balanced data. Model 12 is a LightGBM model trained on the ENN-balanced data.

Models 13 through 16 were trained without using the SMOTE and ENN algorithms for data balancing. The training set consisted of 159 symbiotic stars and 159 non-symbiotic stars. Model 13 was a Decision Tree model that used the Gini index as the splitting criterion. Model 14 was a Decision Tree model that used entropy as the splitting criterion. Model 15 was an XGBoost model, and Model 16 was a LightGBM model.

The optimal model parameters obtained after training are as follows:

Model 1: class_weight = None, criterion = 'gini', max_depth = 7, max_features = 6, min_samples_leaf = 5, min_samples_split = 51, splitter = 'best'.

Model 2: class_weight = None, criterion = 'entropy', max_depth = 10, max_features = 3, min_samples_leaf = 2, min_samples_split = 3, splitter = 'best'.

Model 3: colsample_bytree = 0.6, gamma = 0, learning_rate = 0.1, max_depth = 15, min_child_weight = 1, n_estimators = 580, reg_alpha = 0.1, reg_lambda = 0.1, scale_pos_weight = 300, subsample = 0.9.

Model 4: colsample_bytree = 0.6, learning_rate = 0.1, max_depth = 11, min_child_weight = 1, n_estimators = 500, reg_alpha = 0, reg_lambda = 0, scale_pos_weight = 300, subsample = 0.9.

Model 5: class_weight = None, criterion = 'gini', max_depth = 15, max_features = 5, min_samples_leaf = 2, min_samples_split = 6, splitter = 'best'.

Model 6: class_weight = None, criterion = 'entropy', max_depth = 15, max_features = 7, min_samples_leaf = 2, min_samples_split = 6, splitter = 'best'.

Model 7: colsample_bytree = 0.6, gamma = 0.1, learning_rate = 0.1, max_depth = 15, min_child_weight = 0.1, n_estimators = 100, reg_alpha = 0, reg_lambda = 0, scale_pos_weight = 1, subsample = 0.8.

Model 8: colsample_bytree = 0.6, learning_rate = 0.1, max_depth = 11, min_child_weight = 1, n_estimators = 500, reg_alpha = 0, reg_lambda = 0, scale_pos_weight = 1, subsample = 0.6.

Model 9: class_weight = None, criterion = 'gini', max_depth = 15, max_features = 4, min_samples_leaf = 3, min_samples_split = 6, splitter = 'best'.

Model 10: class_weight = None, criterion = 'entropy', max_depth = 20, max_features = 7, min_samples_leaf = 9, min_samples_split = 15, splitter = 'best'.

Model 11: colsample_bytree = 0.6, gamma = 0.1, learning_rate = 0.1, max_depth = 15, min_child_weight = 0.1, n_estimators = 100, reg_alpha = 0, reg_lambda = 0, scale_pos_weight = 200, subsample = 0.8.

Model 12: colsampe_bytree = 0.6, learning_rate = 0.1, max_depth = 11, min_child_weight = 1, n_estimators = 500, reg_alpha = 0, reg_lambda = 0, scale_pos_weight = 300, subsample = 0.6.

Model 13: class_weight = None, criterion = 'gini', max_depth = 5, max_features = 4, min_samples_leaf = 5, min_samples_split = 51, splitter = 'best'.

Model 14: class_weight = None, criterion = 'entropy', max_depth = 10, max_features = 5, min_samples_leaf = 5, min_samples_split = 5, splitter = 'best'.

Model 15: colsample_bytree = 0.6, gamma = 1, learning_rate = 0.1, max_depth = 15, min_child_weight = 1, n_estimators = 70, reg_alpha = 1, reg_lambda = 1, scale_pos_weight = 1, subsample = 0.9.

Model 16: colsample_bytree = 0.6, learning_rate = 0.1, max_depth = 10, min_child_weight = 1, n_estimators = 50, reg_alpha = 0, reg_lambda = 0, scale_pos_weight = 1, subsample = 0.9.